TecnoLógicas
ISSN-p 0123-7799
ISSN-e 2256-5337
Vol. 22, No. 46, pp 77-97
Sep-dic de 2019

Artículo de Investigación/Research Article

Estimación del potencial fotovoltaico mediante minería de datos en cuatro ciudades de Colombia

Photovoltaic Potential Estimation by Means of Data Mining in Four Colombian Cities

Harrynson Ramírez-Murillo[1], Carlos A. Torres-Pinzón[2] y Edwin F. Forero-García[3]

Recibido: 15 de mayo de 2019
Aceptado: 31 de julio de 2019

Cómo citar / How to cite

H. Ramírez-Murillo, C. A. Torres-Pinzón y E. F. Forero-García, "Estimación del potencial fotovoltaico mediante minería de datos en cuatro ciudades de Colombia", *TecnoLógicas*, vol. 22, no. 46, pp. 77- 97, 2019. https://doi.org/10.22430/22565337.1345

[1] PhD en Ingeniería Electrónica, Facultad de Ingeniería Mecánica, Electrónica y Biomédica (FIMEB), Universidad Antonio Nariño, Manizales-Colombia, harrynson.r@uan.edu.co
[2] PhD en Ingeniería Electrónica, Facultad de Ingeniería Electrónica, Universidad Santo Tomás, Bogotá-Colombia, carlostorresp@usantotomas.edu.co
[3] MSc en Ingeniería Electrónica, Facultad de Ingeniería Electrónica, Universidad Santo Tomás, Bogotá-Colombia, edwinforero@usantotomas.edu.co

Resumen

En este trabajo se analiza el potencial fotovoltaico en cuatro ciudades colombianas, gracias a la información recopilada en Bogotá, Cúcuta, Manizales y Pasto. La metodología propuesta utiliza técnicas de agrupamiento que se implementan mediante el uso del software MATLAB®. Se exponen dos algoritmos de comparación presentados: K-means y Fuzzy C-means, y uno de visualización que es el Análisis de componentes principales (PCA) que ayuda en el análisis de resultados. Este artículo muestra estudios previos relacionados con la minería de datos y se describen los algoritmos mencionados anteriormente. Por otro lado, los resultados y discusión más relevantes, que corresponden a la factibilidad de implementación de las micro-redes, se determinan mediante el cálculo del Factor de Capacidad.

Palabras clave

Algoritmos de clustering, minería de datos, micro-redes, generación de energía solar, sistemas fotovoltaicos.

Abstract

This work analyzes the photovoltaic potential of four cities in Colombia—Bogotá, Cúcuta, Manizales, and Pasto—using information collected in situ and data mining strategies. The methodology of this study is based on clustering techniques implemented in MATLAB® software. Two comparison algorithms are presented: K-means, Fuzzy C-means, and an additional visualization algorithm, i.e., Principal Component Analysis (PCA), which supports results analysis. This article explores published studies regarding data mining and it describes the previously mentioned algorithms. On the other hand, the most relevant results and discussion, which are related to the feasibility of implementation of micro-grids, are determined by calculating the Capacity Factor.

Keywords

Clustering, Plant Factor, Micro-grids, Data Mining, Photovoltaic Potential.

1. INTRODUCCIÓN

Desde la Universidad Santo Tomás y la Universidad Antonio Nariño, ha surgido el interés de abordar una problemática actual: la diversificación del mercado energético colombiano, mediante el estudio de factibilidad de implementar micro redes basadas en energías no convencionales. Por tal motivo, la investigación se enmarcó en el proyecto de convocatoria interna titulado "Evaluación y análisis del potencial energético renovable de las energías eólica y solar en 10 sedes de la Universidad Antonio Nariño" y se obtuvo información recopilada durante un año de funcionamiento de cuatro estaciones meteorológicas en las ciudades de Bogotá, Cúcuta, Manizales y Pasto, adquiriendo datos cada dos minutos, lo que demandó un alto costo computacional para su administración.

En la actualidad, las técnicas para el manejo de grandes volúmenes de información surgen con el nombre de minería de datos y se constituyen como la alternativa ideal en estos casos. Así, desde la Universidad Santo Tomás, se ejecutó el proyecto titulado "Sistemas distribuidos para el control del modo de operación en micro-redes inteligentes basados en técnicas multi-agente", derivándose trabajos de investigación tales como en [1].

El data *mining*, por su nombre en inglés, comprende un conjunto de técnicas tendientes a "realzar" o descubrir comportamientos y patrones presentes dentro de un conjunto de datos o data set, con la finalidad de generar conocimiento que proporcione soluciones a un problema determinado. Dichas técnicas van de la mano con el manejo de un software de cómputo, sin el cual no sería posible su desarrollo, y que permite deducir que efectivamente su aparición dentro de las ciencias estadísticas no solo es muy nueva, sino que está cambiando de forma constante y aún no están definidas completamente [2].

Ante la perspectiva futura del agotamiento de los recursos tradicionales energéticos, las fuentes de energía no convencionales han venido tomando fuerza y buscan un espacio dentro de la canasta de oferta energética. Diferentes acuerdos internacionales y muchas políticas internas en varios países apuntan en esta dirección. Colombia no puede ser ajena a dicha coyuntura y ha desarrollado estrategias para afrontar los nuevos retos.

A la luz de lo anterior, se evidencia que existe un favorecimiento para la implementación de energías renovables y, de esta forma, poder aportar a la protección del medio ambiente, mejorando la calidad del servicio sin riesgo de apagones como los sucedidos durante los años de 1992 y 2015 a causa del fenómeno del niño, cuando se generó una reducción en los niveles de los embalses y cauces en los ríos, afectando la producción de energía hidráulica, situación por la cual no se pudo cubrir con la demanda total de energía del país, perturbando en forma drástica la economía nacional [3].

Por lo tanto, se presenta una oportunidad ya que los paneles solares han disminuido de costo en comparación con años anteriores, logrando de esta manera incursionar cada vez más en la canasta energética. Así mismo, el gobierno nacional proporciona incentivos, como los establecidos en Ley 1715 de 2014, donde se establecen estímulos a la inversión en proyectos de Fuentes No Convencionales de Energías Renovables (FNCER), disminuyendo el impuesto de renta y complementarios [4].

En los últimos años diversas técnicas de *data mining* se vienen aplicando exitosamente en varias áreas del conocimiento, desde investigaciones médicas hasta estudios de mercados [5]. En el campo de la ingeniería no hay muchas publicaciones que usen la minería de datos con el objeto de aprovechar la energía proveniente de la radiación solar y es allí donde se enfoca el presente documento.

Mediante este trabajo de investigación, se pretende procesar y caracterizar la

información obtenida de una red de estaciones meteorológicas, conformadas por las variables radiación solar (Rad) en W/m², temperatura (T) en °C y hora (h) en h, ubicadas en cuatro (4) ciudades de Colombia, con la finalidad de determinar la viabilidad de implementar micro-redes, basadas en energía solar fotovoltaica, a través del cálculo del Factor de planta. Para ello, se mencionan los antecedentes más relevantes, donde se incluye el correspondiente marco regulatorio colombiano, experiencia en estudios y minería de datos, aplicados a radiación solar. Por otro lado, en la sección de Materiales y métodos, se mencionan las características de las estaciones meteorológicas usadas, junto a los diferentes algoritmos considerados: K-means y Fuzzy C-means, y uno de visualización que es el Análisis de componentes principales (PCA). Es importante aclarar que, para la ejecución de dichos algoritmos, mencionados anteriormente, se hace uso de la herramienta Fuzzy Clustering and Data Analysis Toolbox [6], con la ayuda del software de MATLAB®. Finalmente, se ilustran los diferentes resultados, en los cuales se establece la factibilidad de implementar sistemas fotovoltaicos, para cada una de las regiones objeto de estudio, junto a las correspondientes conclusiones.

2. ANTECEDENTES

2.1 Marco regulatorio colombiano

Gracias a la estrategia de Pagos por Servicios Ambientales (PSA), implementada por el gobierno colombiano, se han dado grandes avances dentro de los esfuerzos tendientes a reducir la huella de carbono, siendo uno de los más claros ejemplos de PSA el BanCO_2 que ha beneficiado a 2.600 familias y tiene casi 26 mil hectáreas conservadas [7]. De igual manera, dentro de dicho marco legal de políticas de desarrollo sostenible, se destacan los estímulos otorgados por la ley 142 y 143 de 1994, la llamada ley URE (Uso Racional de Energía) de 2001 y la más actual, y tal vez la más concerniente a este trabajo, la Ley 1715 de 2014 que regula la integración de las energías renovables no convencionales al sistema energético nacional. Dentro de dichas políticas, el Estado ha buscado fomentar la inversión privada procedente principalmente de Europa, Estados Unidos y en los últimos años de China, conformando tres grandes conglomerados generadores de energía (EPM, EMGENSA e ISAGEN), con grandes proyectos de interconexión en proceso como Ituango y Porvenir II, que buscan una línea de transmisión que entrelace el país con Centroamérica (Colombia-Panamá), con una capacidad de 100 MW y posteriormente con Ecuador con una capacidad de 300 MW. En la actualidad se están llevando a cabo otros proyectos, tal como se indican en la Tabla 1.

Tabla 1. Proyectos energéticos en desarrollo actualmente en Colombia. Fuente: [8]

Tipo de proyecto	Potencia (MW)
97 hidráulicos	3631
8 térmicos	858
4 eólicos	654
1 solar	19.9

2.2 Experiencia en estudios de radiación

Los logros más significativos provienen entonces en su mayoría de investigadores particulares y/o entidades relacionadas con la academia. Se ha querido entonces estudiar el fenómeno de la radiación como fuente primaria para construir células fotoeléctricas. En Latinoamérica son más conocidos los estimativos a partir de la formulación de Suerhcke que fueron usados por Hernández en Uruguay [9], los trabajos de Grossi y Roberti en Argentina [10], y antes los de Huertas en México [11]. En el caso de Colombia, Rodríguez Murcia [12] presenta un recuento interesante del historial de uso de la energía solar en

nuestro país, aportando además las perspectivas de desarrollo de la misma.

Mientras unos autores buscan predecir la cantidad de radiación en el corto y mediano plazo, como Zamarbide [13], otros proponen la construcción de solarímetros caseros [14] que, como su nombre lo indica, son equipos sencillos con materiales de bajo costo y fácil adquisición, que se muestran como alternativas a complejas relaciones matemáticas, como las presentadas hace años por respetados autores en la materia que buscan relaciones entre los ángulos que describen la posición del sol [15].

En Colombia se han planteado estrategias de optimización, basadas en la redistribución de energía generada, evitando consumirla en los momentos en los cuales las tarifas de energía son mayores, haciendo uso de un sistema de almacenamiento. Para ello, se realiza un análisis económico, basado en un flujo de fondos neto, para un periodo de 20 años, dado que ese es considerado el tiempo de vida útil de los paneles solares [16]. Así mismo, otros autores proponen el modelado de sistemas fotovoltaicos, considerando su comportamiento, tanto en el modo de operación directo, como en modo inverso bajo condiciones no uniformes de irradiación. Este modelo matemático permite evaluar la eficiencia de un sistema fotovoltaico, en tiempos cortos de simulación, para sistemas de gran tamaño [17]. También se han realizado estudios donde se modelan sistemas fotovoltaicos, operando bajo condiciones tanto regulares como irregulares, donde se presentan mejoras en exactitud y tiempos de cómputo en evaluaciones energéticas, que son de gran utilidad en el diseño de plantas fotovoltaicas [18].

2.3 Antecedentes en estudios de minería de datos aplicada en radiación solar

Es importante mencionar que el KDD (Knowledge Discovery in Data Bases) es una metodología relativamente nueva y que la minería de datos, a pesar de su enorme utilidad, solo la ha implementado en los últimos años y su desarrollo, como se dijo en apartados anteriores, está estrechamente ligado con los avances en computación. Muchas publicaciones corresponden a trabajos de grado, para maestrías doctorados, en el campo de las ingenierías eléctrica y electrónica. Así, por ejemplo, se tienen los análisis estadísticos con predicción de series temporales de Pomarés en 2012 [19].

En el país la implementación de proyectos de generación de energía con fuentes no convencionales es incipiente. No obstante, el Instituto de Planificación y Promoción de Soluciones Energéticas, para Zonas No Interconectadas (IPSE) ha venido impulsando el uso de energías renovables en las localidades apartadas sin servicio de energía. La experiencia forjada en este ejercicio faculta al instituto para liderar la reconversión de diversificación del cambio de la matriz energética en Colombia [20].

En Colombia los estudios predictivos usando *Big Data* no han sido muy divulgados y las cuestiones relacionadas con sistemas fotovoltaicos se han centrado más en su uso directo, gracias a la disponibilidad de gran cantidad de radiación solar en casi todo el territorio nacional por su posición ecuatorial en el globo terráqueo. Ciudades de la Costa Atlántica han sido objetivo directo de las compañías proveedoras de fuentes solares como Tecnoglass que ha centrado esfuerzos en Barranquilla [21].

3 MATERIALES Y MÉTODOS

3.1 Estación meteorológica empleada

La estación meteorológica PVMET-300, indicada en la Fig. 1, se encuentra diseñada para cumplir con ciertas necesidades requeridas por las energías alternativas, especialmente las derivadas del sol, ya que cuenta con los elementos y las características descritas a continuación,

muy útiles, a pesar de su forma compacta, lo que conlleva a una instalación sencilla en diferentes lugares con difícil acceso:

-Sensor de temperatura ambiente
-Sensor de velocidad del viento
-Sensor de dirección del viento
-Sensor de humedad relativa
-Sensores de temperatura en la parte posterior del PV
-Comunicación inalámbrica con el registrador de datos Modbus
-Comunicación Modbus RS-485
-Sensor global de irradiación solar
-Sensor de presión barométrica
-Pluviómetro

3.2 Índices de validación

Es importante evaluar el resultado de los algoritmos de clustering, pero es difícil definir cuándo el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado [22]. La validación se refiere a si en determinada partición encajan bien los datos. Aunque los algoritmos de clustering siempre tratan de hallar el mejor ajuste dentro de los parámetros establecidos, esto no significa necesariamente que el valor que hallen sea el adecuado o el mejor [23]. En muchos casos el criterio del investigador puede jugar un papel muy importante al respecto.

3.2.1 *Índice de silueta (S(i))*

Este índice trabaja de la siguiente manera: para cada punto x dentro de un grupo, primero calcula el promedio de la distancia entre este y todos los otros puntos dentro del mismo clúster (cohesión). Luego, evalúa el promedio de la distancia entre este mismo punto x y todos los otros puntos del clúster más cercano (separación). El coeficiente de silueta para el punto x está definido como la diferencia entre la separación y la cohesión dividida por el más grande de los dos [24]. Es de anotar que esta distancia se puede hallar por cualquiera de los métodos conocidos como puede ser Euclídea, Manhattan, Mahalanobis, etc. El procedimiento se repite con cada punto de cada clúster. El número óptimo de clústeres corresponde a su valor máximo.

Fig. 1. Estación meteorológica PVMET-300 utilizada en la toma de datos.
Fuente: autores.

3.2.2 Índice de partición (SC)

Se define como la relación o el cociente de la sumatoria de todas las cohesiones entre la sumatoria de las separaciones del clúster. Es una suma de medidas individuales de validez de los grupos, normalizada a través de la división por la cardinalidad difusa de cada grupo. El número óptimo de clústeres se encuentra donde esté mayor "caída" o descenso de los datos en la gráfica correspondiente [25].

3.2.3 Índice de separación (SI)

Contrario al índice de partición (SC), el índice de separación (SI) usa la distancia mínima de separación para validar la partición [23]. Su interpretación es igual al índice de partición (SC) encontrando el número óptimo de clúster en el punto donde los datos presenten el mayor descenso.

3.2.4 Índice de Xie y Beni (XB)

Su propósito consiste en cuantificar la proporción de la variación total dentro del clúster, respecto a la separación de los mismos [26]. Dicho de otra forma, se define como el cociente entre la varianza total y la mínima separación de los clusters. El número óptimo de clústeres se encuentra a través del primer mínimo local en la gráfica.

3.3 Algoritmo K-means

K-means es un método de clustering (descriptivo) de tipo no jerárquico y no supervisado; es decir, no requiere un conocimiento previo de los datos a analizar, ajustándose a lo que entregue la fuente primaria de Big Data, sin juicios a priori de los mismos, como se muestra en la Fig. 2, y que necesita de las ecuaciones (1) y (2). Estos métodos de partición son populares por su simplicidad y porque precisan de poca capacidad de procesamiento por parte del equipo de cómputo usado, comparado con otros métodos como Redes Neuronales Artificiales (RNA) y Maquinas de Soporte Vectorial (MSV). Sin embargo, para garantizar la efectividad de estos algoritmos se hace necesario que el proceso sea iterativo, de tal forma que el método "aprenda" minimizando así el error de la distancia euclídea, de todos los objetos a los k centroides, permitiendo que sea lo más acertada posible, dentro del clúster A_i.

$$J(X,V) = \sum_{i=1}^{c} \sum_{k \in A_i} \|X_k - V_i\|^2 \quad (1)$$

$$J(X,V) = \sum_{i=1}^{c} \sum_{k \in A_i} \|X_k - V_i\|^2 \quad (2)$$

"La eficacia del algoritmo K-means depende de la idoneidad del parámetro K. Si este es mayor o menor que el número real de grupos, se crean grupos ficticios o se agrupan objetos que deberán pertenecer a clúster distinto" [27], hecho por el cual se requiere de un conocimiento a priori del número óptimo de conglomerados de datos c.

donde:

K: número de clúster considerados inicialmente
i: Clúster comprendido entre 1 y K
c: Valor óptimo de particiones
V_i: Centroide de cada clúster
J: Función de costo
A_i: Datos contenidos en el clúster i, cuyo centroide corresponde a vi
$iter$: Número de iteraciones
N: Número de objetos en el conjunto de datos

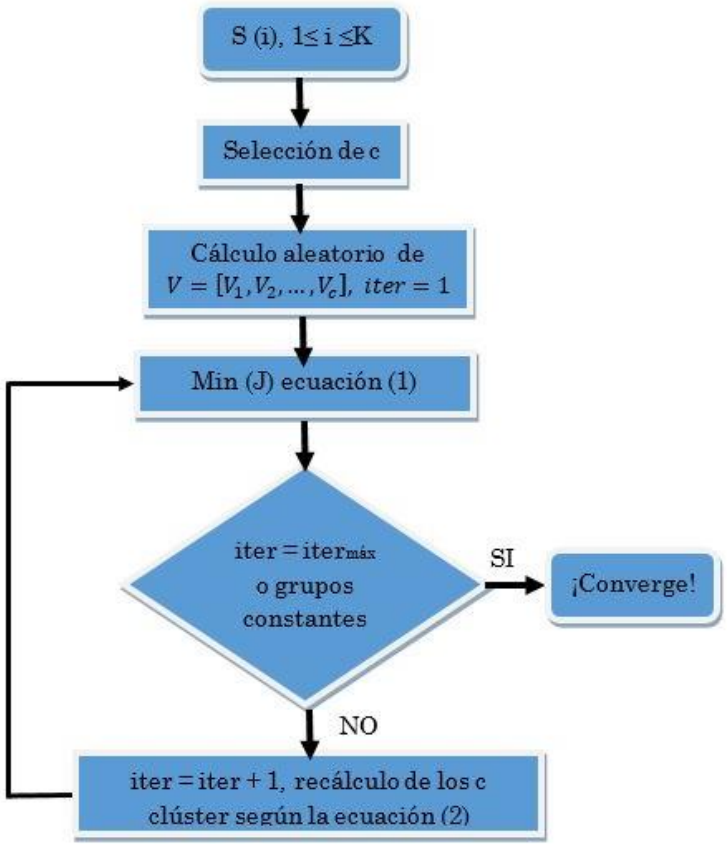

Fig. 2. Representación gráfica del algoritmo k-medias. Fuente: autores.

3.4 Algoritmo K-means

El Fuzzy C-means es uno de los algoritmos de clustering de partición difusa más conocido y muy utilizado en el campo de la minería de datos. Encuentra uso en múltiples aplicaciones como es el caso del procesado de imágenes para el envío por redes telemáticas y diferentes aplicaciones a nivel de ingeniería. Este algoritmo se basa en la realización de iteraciones, fijando de manera aleatoria las coordenadas de los centroides y calculando tanto la distancia como también si pertenece o no cada dato a un determinado clúster. Para ello Fuzzy C-means toma la distancia euclídea menor de un dato, asignándolo al clúster más cercano y posteriormente recalcula los centroides en cada iteración hasta llegar al punto de convergencia de los datos. En este punto se tiene el criterio para detener el proceso iterativo [28]. Este algoritmo, que requiere de las ecuaciones (4), (5) y (6), se resume en el diagrama de flujo de la Fig. 3.

El algoritmo Fuzzy C-means se fundamenta en la minimización de una función objetivo llamada *J* como lo indica la ecuación (3), que es la base de este algoritmo, pues "la mayoría de las técnicas analíticas de fuzzy clustering se basan en la optimización de la función objetivo C-means" [28] o en algunos casos una modificación de esta, como por ejemplo, los algoritmos que adaptan una norma distinta para cada clúster, permitiendo detectar tamaños y formas diferentes como es el caso de Gustafson-Kessel (GK).

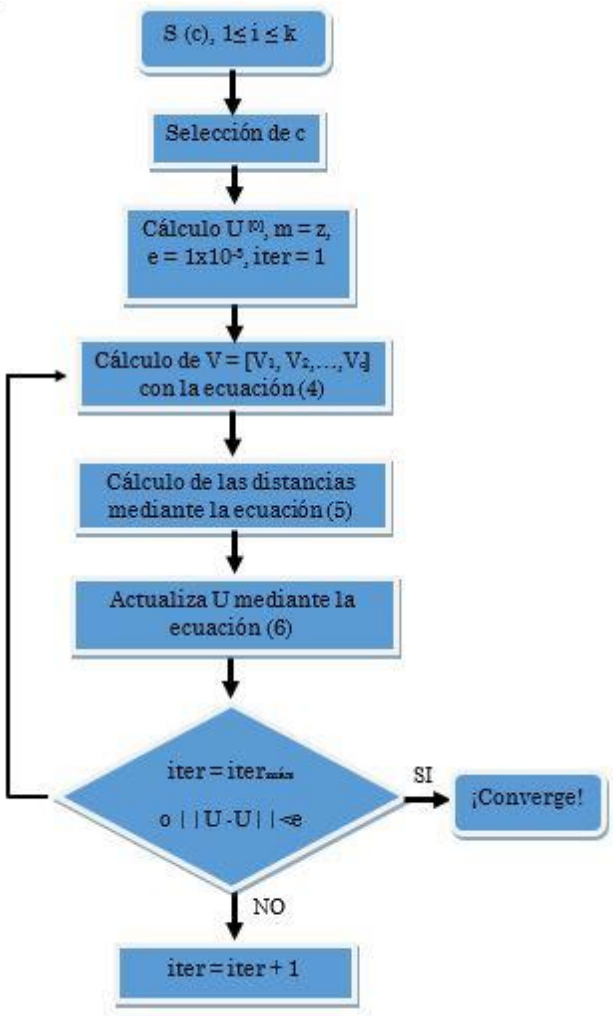

Fig. 3. Diagrama de flujo del algoritmo Fuzzy C-means. Fuente: autores.

$$J(X,U,V) = \sum_{i=1}^{c}\sum_{k\in A_i}(\mu_{ik})^m \|X_k - V_i\|^2 \qquad (3)$$

Las coordenadas de prototipos de cluster $V_{i(1)}$ vienen dadas por:

$$V_i^{(l)} = \frac{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m\ X_k}{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m}, 1 \le i \le c \qquad (4)$$

Las distancias se calculan bajo la norma Euclídea así:

$$D_{ikA}^2 = (X_k - V_i)^T\ A\ (X_K - V_i), \quad 1 \le i \le c, \quad 1 \le k \le N \qquad (5)$$

La matriz de partición U es:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}(\frac{D_{ikA}}{D_{jkA}})^{\frac{2}{(m-1)}}} \qquad (6)$$

donde, A corresponde a la inversa de la matriz de covarianza o de dispersión de datos F (A= F-1) y se muestra en la ecuación (7):

$$F = \frac{1}{N}\sum_{K=1}^{N}(X_K - \bar{X}_K)(X_K - \bar{X}_K)^T \qquad (7)$$

3.5 Análisis de componentes principales (PCA)

El Análisis de componentes principales consiste en encontrar transformación ortogonal de las variables

originales para conseguir un nuevo conjunto de variables incorreladas, denominadas Componentes principales, que se obtienen en orden decreciente de importancia [29].

El PCA se usa básicamente como un elemento de visualización de los datos, permitiendo pasar de un espacio vectorial definido en R^3 a un subespacio definido en R^2, es decir aplicar una transformación lineal $T: R^3 \rightarrow R^2$, sin pérdida alguna de información relevante, maximizando su varianza, donde se consideran las dos componentes principales de "radiación" y "temperatura" en R^3. Dicha proyección se realiza gracias al uso del análisis de "autovalores", también conocidos como "valores propios". El autovector asociado con el autovalor más grande tiene la misma dirección del primer componente principal, y el autovector asociado con el segundo autovalor más grande, determina la dirección del segundo componente principal. PCA calcula la matriz de auto correlación de los datos con sus vectores propios, ordenándolos de acuerdo a sus valores propios y luego son normalizados. El sistema de mapeo usa los dos primeros autovalores no nulos, junto a sus correspondientes autovectores de la matriz de covarianza F según la ecuación (8).

$$F = P\Lambda_i P^T \quad (8)$$

donde:

Λ es la matriz que incluye los valores propios λii de la diagonal de F en orden decreciente.

P es la matriz que incluye los vectores propios.

En el caso de estudio de esta investigación se ha considerado $Z = X - \underline{X}$ como la matriz de N muestras centradas pertenecientes a un espacio vectorial en R^3, en el cual tenemos las variables "radiación" (Rad), "temperatura" (T), y "hora" (h) al cual se le realiza PCA, para proyectarlo a un subespacio de dimensión 2, el cual se va a llamar plano L, obteniéndose la matriz Y, de tal forma que la proyección ortogonal de las muestras sobre L maximice su varianza y permita el estudio de las correlaciones entre las componentes de Z, clúster, centroides, entre otras, sin pérdida alguna de información relevante.

donde:

$Y = W^{-1}Z$, corresponde a la representación bidimensional de la matriz de observaciones centradas Z.

$W = P_{3x2}\Lambda_{2x2}^{1/2}$, corresponde a la matriz de valores de peso que contiene 2 componentes principales en sus columnas.

P_{3x2} se encuentra conformada por las dos primeras columnas de P, que corresponden a los vectores propios más representativos.

Λ_{2x2} es una matriz diagonal con los dos primeros valores propios.

3.6 Regresión polinómica

Al realizar el correspondiente ajuste de los datos de las variables continúas "radiación" y "hora" con el proceso de clustering [29], se pueden ajustar mediante una regresión polinómica, obteniéndose la ecuación (9):

$$Rad(h) = ah^2 + bh + c \quad (9)$$

Una vez obtenidos los coeficientes a, b y c de la ecuación, junto con el coeficiente de correlación r^2, que permite verificar la fiabilidad del ajuste polinómico realizado, es posible estimar la cantidad de energía disponible diaria por metro cuadrado, o mejor conocida como irradiación I, tal como se muestra en la ecuación (10).

$$I = \int_{h_1}^{h_2} Rad(h)dh \ (J/m^2) \quad (10)$$

donde, h_1 y h_2 corresponden a los puntos de corte sobre el eje de las abscisas, es decir sobre el eje h. Estos parámetros corresponden a las horas de salida y caída del sol, estimadas para cada una de las ciudades objeto de estudio.

3.7 Factor de planta (Fp)

También conocido como factor de carga o de capacidad, es la fracción porcentual entre la energía producida y la nominal, para un determinado período de tiempo. Depende de la presencia del recurso en un determinado período y de la tecnología utilizada. Típicamente toma valores entre el 20-50% [27]. Su expresión en porcentaje viene dada por la ecuación (11):

$$Fp(\%) = \frac{(P_p)(t)(h_2 - h_1)}{(365)(a_p)(b_p)I} * 100 \quad (11)$$

Donde, si se considera un panel mono cristalino de 100 W comercial, se tiene que:

P_p: capacidad del panel (100 W)
t: tiempo de estudio (365 días)
h_2 y h_1: raíces de la regresión polinomial obtenidas mediante el comando de MATLAB® polyfit
a_p: ancho del panel (0.67 m)
b_p: largo del panel (1.17 m)
I: irradiación diaria estimada (J/m^2)

4. RESULTADOS Y DISCUSIÓN

Dentro de las principales funciones de los algoritmos considerados en este estudio se encuentran:

-Establecer el número óptimo de conglomerados mediante el uso de índices de validación.
-Determinar las coordenadas de cada uno de los centroides para las diferentes ciudades objeto de estudio.
-Calcular el correspondiente Factor de planta para establecer la factibilidad de implementar sistemas fotovoltaicos.
-Estimar los tiempos de cómputo promedio e iteraciones para la ejecución de cada algoritmo, en los casos de K-means y Fuzzy C-means.
-Visualizar las variables objetos de estudio (Rad, T, h), en un espacio de dimensión reducida, sin pérdida de información, mediante el uso de PCA.

Para establecer el número de clústeres óptimo, en cada algoritmo se determinan a partir de los índices de validación mencionados en la Sección 3.2. A manera de ejemplo, se muestran los resultados obtenidos para cada estación meteorológica en la Fig. 4, donde se calcula el Índice de Silueta $S(i)$.

Al considerarse el número óptimo de clústeres a partir del $S(i)$ como su valor máximo para las sedes de Bogotá, Cúcuta y Manizales respectivamente, se obtiene un valor igual a 3, tal como lo indica la Fig. 4 (a), 4(b) y 4(c), pero para la ciudad de Pasto se nota una diferencia, obteniéndose como resultado un valor de 2 como el óptimo, tal como se muestra en Fig. 4 (d). Estos resultados son comparados posteriormente con los otros índices antes mencionados. De esta forma, ha sido seleccionado el número óptimo de conglomerados como un valor de 3, cuyas etiquetas corresponden a "mañana", "medio día" y "tarde" para cada uno de ellos. Es importante anotar que dichas figuras se encuentran sujetas a un proceso de inicialización aleatorio; no obstante, el número óptimo de centroides elegido se mantiene sujeto al criterio del investigador.

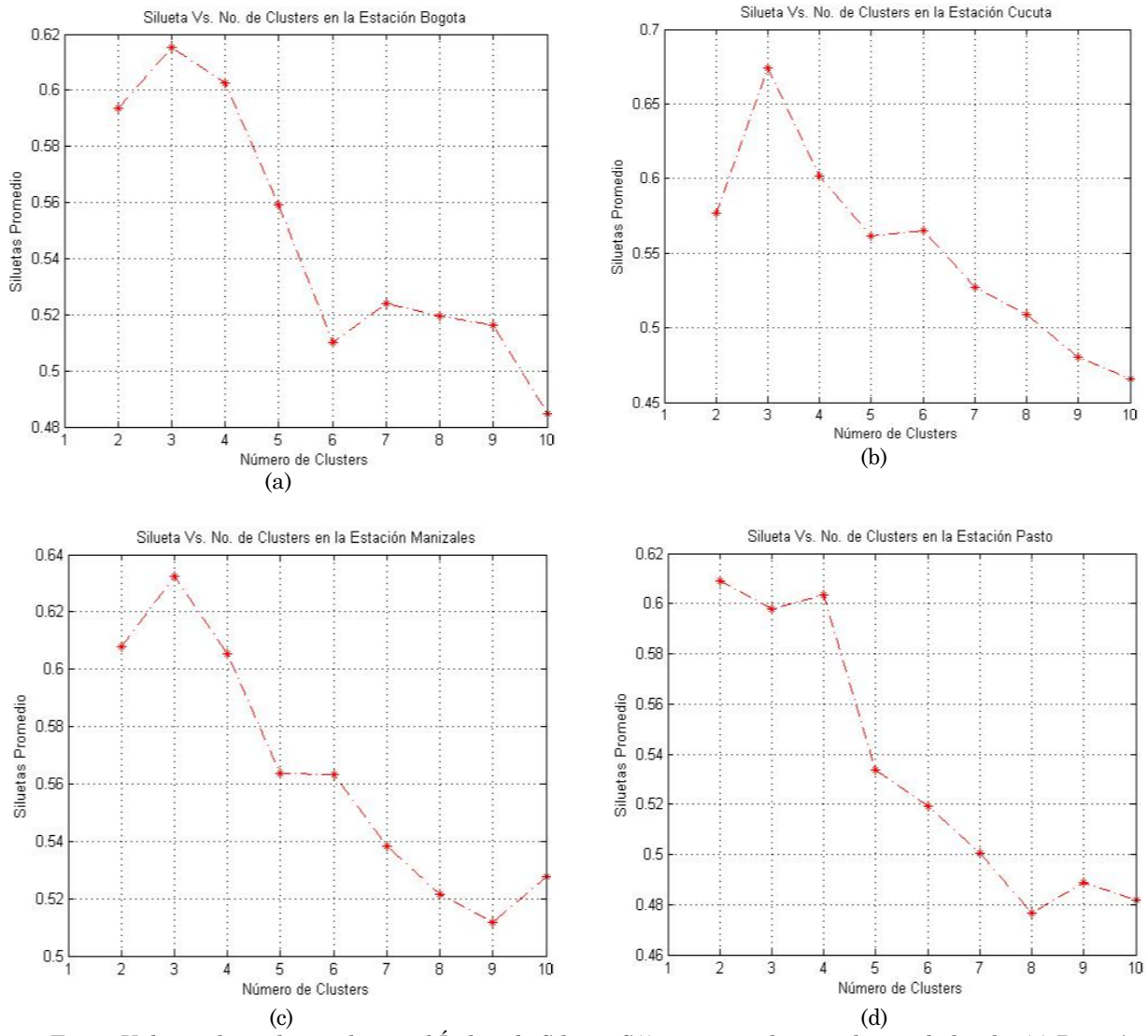

Fig. 4. Valores obtenidos mediante el Índice de Silueta S(i) correspondiente a las ciudades de: (a) Bogotá, (b) Cúcuta, (c) Manizales y (d) Pasto, utilizando el algoritmo k-means. Fuente: autores

4.1 Algoritmo K-means

Dentro de las modificaciones realizadas al Toolbox, el algoritmo muestra una gráfica en R^3, tal como lo indican las Fig. 5(a), 5(b), 5(c) y 5(d), en las cuales se pueden observar las clasificaciones de los tres grupos de conglomerados, obtenidos mediante el análisis de índices de validación descritos anteriormente, graficando en sus variables normalizadas: hora, temperatura y radiación. Las coordenadas de cada uno de sus centroides se muestran en la Tabla 2.

Así como también se obtiene una gráfica proyectada en R^2, sin perder información como se muestra en las Fig. 6(a), 6(b), 6(c) y 6(d), en las cuales se observan curvas de nivel elípticas, debido a la interacción existente entre los conglomerados, cada isolínea se representa por un color diferente que indica un valor de proximidad a cada clúster, correspondiendo a 0.9 el color que más pertenece al conjunto y 0.4 el más lejano al centroide.

Posteriormente, se desnormalizan los datos mediante la ecuación (12) y se presentan en la Tabla 3.

Tabla 2. Coordenadas de los centroides de la gráfica R3 para cada una de las estaciones metereologicas. Variables normalizadas, empleando el método de agrupamiento K-Means. Fuente: autores.

Variable normalizada	Hora	Temperatura	Radiación
Bogotá	0.1616	0.3524	0.1199
	0.4825	0.6199	0.4232
	0.8203	0.5361	0.1274
Cúcuta	0.1709	0.3144	0.1356
	0.4924	0.6261	0.4915
	0.8255	0.5047	0.1219
Manizales	0.1730	0.2812	0.1312
	0.4939	0.5812	0.4433
	0.8140	0.5143	0.1282
Pasto	0.1683	0.3201	0.1069
	0.4924	0.5516	0.3827
	0.8242	0.4800	0.1237

Tabla 3. Coordenadas de los centroides de la gráfica R^3 para cada una de las estaciones metereologicas. Variables desnormalizadas, trabajando con el método de agrupamiento K-Means. Fuente: autores.

Variable real	Hora (h)	Temperatura (°C)	Radiación (W/m²)
Bogotá	7.9360	14.4896	170.3352
	11.7817	19.4925	601.3252
	15.8298	17.9257	180.9779
Cúcuta	8.0447	27.4256	191.3383
	11.8928	33.5976	693.4469
	15.8786	31.1939	171.9764
Manizales	8.0701	17.3793	182.8628
	11.9097	22.4792	617.8958
	15.7406	21.3426	178.7721
Pasto	8,0138	13.2212	157.4167
	11.8927	16.9258	563.7766
	15.8620	15.7796	182.2821

$$X_{jl} = X_{jlnom}\left[\left((X_l) - min(X_l)\right)\right] + min(X_l)\,; \quad \text{j=1,2,...,N; l= 1, 2, 3} \tag{12}$$

Con esta información se puede determinar el área bajo la curva en el diagrama de Rad Vs. h, permitiendo conocer de esta forma la energía disponible en cada estación, de acuerdo con la ecuación (10), obteniéndose de esta forma el Factor de planta como se muestra en la Tabla 4. Para este caso se utiliza la información correspondiente a la de un panel fotovoltaico de la Sección 3.7.

Para evaluar el desempeño de los algoritmos estudiados, evitando así

tiempos muy prolongados de cómputo, se han realizado 100 simulaciones, inicializando aleatoriamente cada una de ellas, obteniéndose los mismos centroides en todos los casos, pero con diferentes tiempos de convergencia, lo cual se encuentra estrechamente relacionado con el número de iteraciones y está en las Tablas 5 y 6.

Tabla 4. Factor de planta para cada una de las ciudades mediante el método de K-means. Fuente: autores.

Ciudad	Factor de planta (%)
Bogotá	31.7980
Manizales	27.5928
Cúcuta	30.9697
Pasto	33.9286

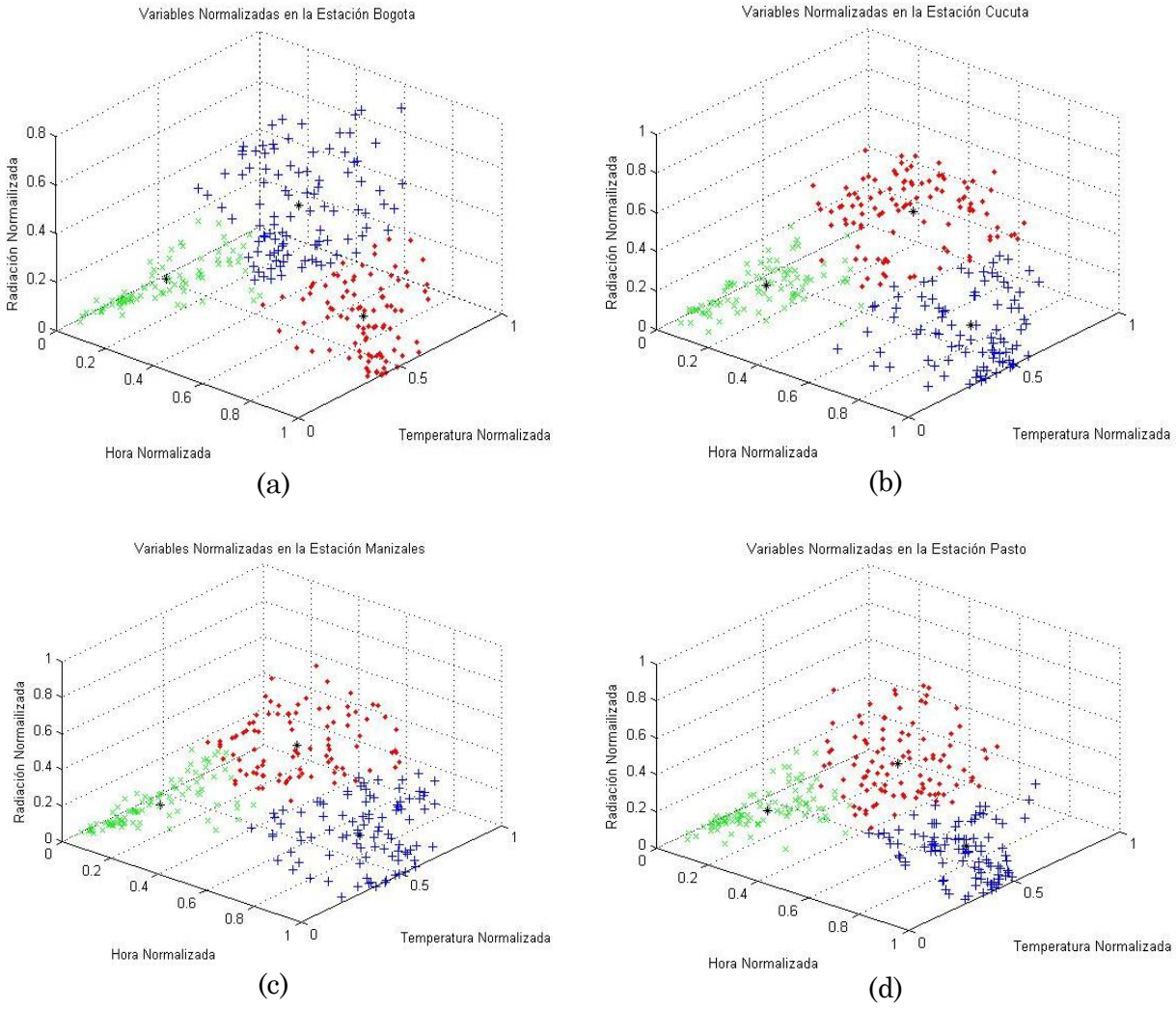

Fig. 5. Variables normalizadas en el espacio R^3, correspondiente a las ciudades de: (a) Bogotá, (b) Cúcuta, (c). Fuente: autores.

Tabla 5. tiempo promedio en segundos requerido de cómputo para realizar los cálculos de las estaciones meteorológicas utilizando el método de K-means. Fuente: autores.

Bogotá	Cúcuta	Manizales	Pasto	Total
15.2528	12.4729	15.4509	15.4208	58.5973

Tabla 6. Parámetros de distribución triangular de iteraciones para cada una de las estaciones meteorológicas con parámetros A, B y C, que corresponden a los valores mínimos, moda y máximo, respectivamente, utilizando el método de agrupamiento k-means. Fuente: autores.

Bogotá			Cúcuta			Manizales			Pasto		
a	b	c	a	b	c	a	b	c	a	b	c
16	27	32	13	18	25	15	24	29	16	25	28

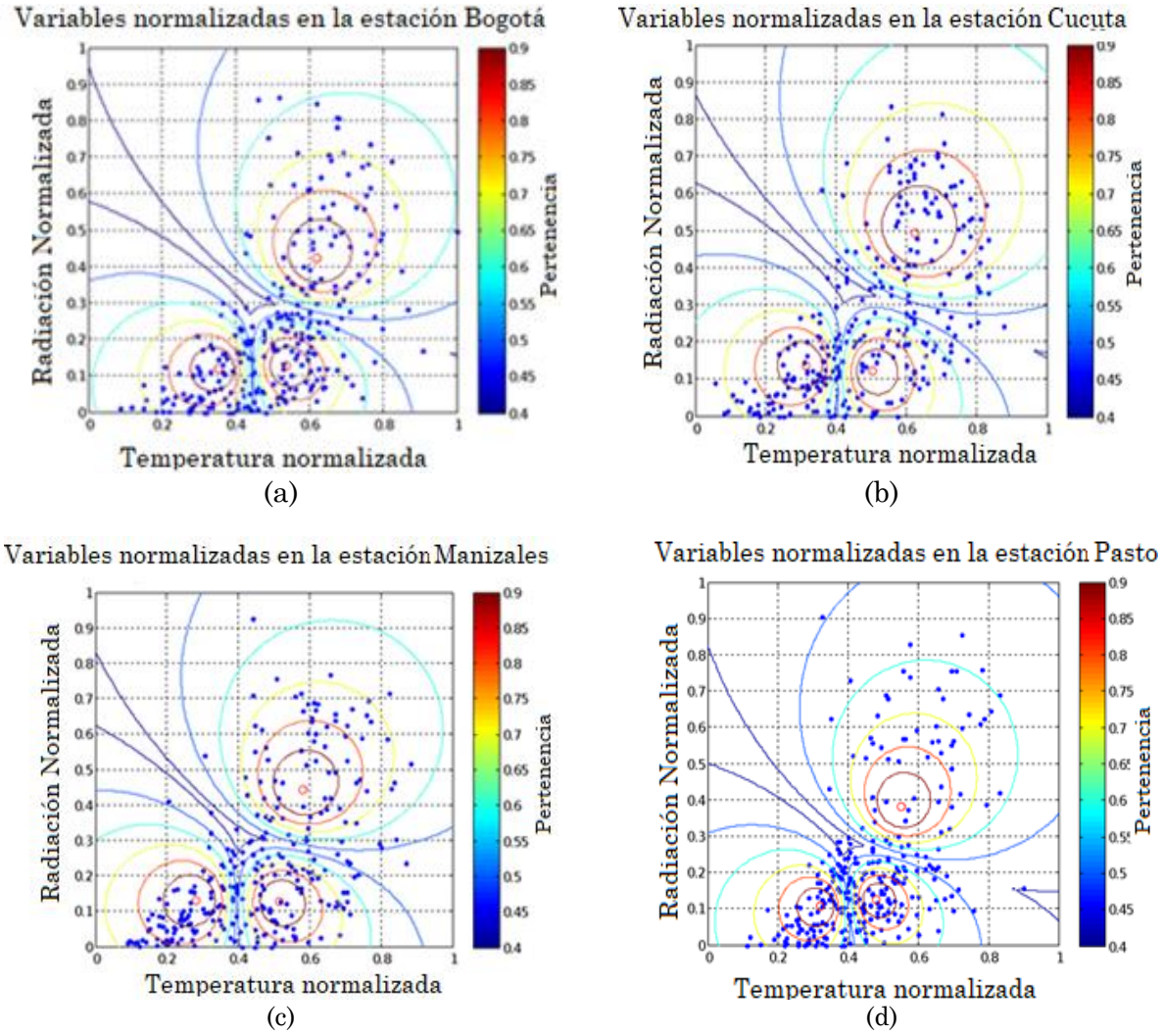

Fig. 6. Variables normalizadas en el plano de Radiación vs Temperatura, correspondiente a las ciudades de: (a) Bogotá, (b) Cúcuta, (c) Manizales y (d) Pasto, utilizando el algoritmo k-means. Fuente: autores.

4.2 Algoritmo Fuzzy C-means

Las coordenadas normalizadas obtenidas a través de este método, para cada uno de sus centroides, se muestran en la Tabla 7. Adicionalmente, se desnormalizan los datos mediante la ecuación (12) y se presentan en la Tabla 8.

De forma similar, se determina el área bajo la curva en el diagrama de Rad Vs. h, permitiendo conocer la energía disponible en cada estación, según la ecuación (10), obteniéndose de esta forma el Factor de planta como se muestra en la Tabla 9.

Para evaluar el desempeño del algoritmo, se determinan los diferentes tiempos de convergencia, lo cual está estrechamente relacionado con el número de iteraciones y se encuentran en las Tablas 10 y 11.

Tabla 7. Coordenada de los centroides de la gráfica R^3 para cada una de las estaciones metereologicas. Variables normalizadas utilizando el método de agrupamiento Fuzzy C means. Fuente: autores.

Variable normalizada	Hora	Temperatura	Radiación
Bogotá	0.1527	0.3431	0.1134
	0.4847	0.6189	0.4316
	0.8243	0.5332	0.1257
Cúcuta	0.1584	0.3058	0.1248
	0.4959	0.6282	0.5046
	0.8378	0.5025	0.1141
Manizales	0.1549	0.2632	0.1152
	0.4880	0.5735	0.4410
	0.8227	0.2540	0.1289
Pasto	0.1567	0.3149	0.1008
	0.4933	0.5433	0.3633
	0.8303	0.4782	0.1244

Tabla 8. Coordenada de los centroides de la gráfica R^3 para cada una de las estaciones meteorológicas. Variables desnormalizadas empleando el algoritmo de Fuzzy C means. Fuente: autores.

Variable normalizada	Hora (h)	Temperatura (°C)	Radiación (W/m^2)
Bogotá	7.8304	14.3159	161.1439
	11.8083	19.7431	613.3427
	15.8783	17.8711	178.5830
Cúcuta	7.8959	27.2558	176.1276
	11.9348	33.6392	711.9874
	16.0258	31.1492	160.9721
Manizales	7.8532	17.0752	160.5654
	11.8394	22.3497	614.6968
	15.8448	21.4465	179.7412
Pasto	7.8747	13.1376	148.4987
	11.9037	16.7928	535.1701
	15.9363	15.7514	183.3070

A partir de los resultados obtenidos para los tiempos de cómputo promedio de las tablas 10 y 11, se aplica este algoritmo con los obtenidos mediante el de Fuzzy C-means, dado que sus tiempos de simulación son inferiores. Las coordenadas normalizadas proyectadas en R^2 obtenidas para las diferentes sedes, se ilustran en la Fig. 7, indicando la pertenencia de los datos a cada uno de los centroides mediante una barra de colores.

Como puede observarse en la Tabla 12, los valores obtenidos para los centroides desnormalizados son similares a los mostrados en la Tabla 8, que corresponden al algoritmo de Fuzzy C-means. Las diferencias entre dichos resultados son debidas a la reducción de espacio mediante el algoritmo de PCA, presentándose un error relativo de reconstrucción, tal como se muestra en la Tabla 13.

Tabla 9. Factor de planta para cada estación, trabajando con el método agrupamiento Fuzzy C Means. Fuente: autores.

Ciudad	Factor de planta (%)
Bogotá	31.1878
Manizales	26.8756
Cúcuta	31.1254
Pasto	35.7410

Tabla 10. Tiempo promedio, en segundos, requerido de cómputo para realizar cálculo de las estaciones meteorológicas, Fuzzy C Means. Fuente: autores.

Bogotá	Cúcuta	Manizales	Pasto	Total
10.4067	10.6936	10.5210	10.9631	42.5844

Tabla 11. Interacciones para cada una de las estaciones meteorológicas de UAN utilizando el método de agrupamiento Fuzzy Means. Fuente: autores.

Bogotá	Cúcuta	Manizales	Pasto
18	13	17	18

Tabla 12. Coordenadas denormalizadas V en R^3 para cada una de las estaciones meteorológicas mediante algoritmo de PCA. Fuente: autores.

Variable real	Temperatura (°C)	Radiación (W/m^2)	Hora (h)
	19.5690	608.7867	11.7909
Bogotá	14.3475	159.6380	7.8243
	17.9064	176.9768	15.8725
	33.9537	696.8140	11.8740
Cúcuta	27.0304	186.9930	7.9394
	31.0926	163.6978	16.0367
	22.5775	601.9542	11.7748
Manizales	16.8858	170.8883	7.9047
	21.4099	181.8385	15.8551
	16.7753	535.9163	11.9061
Pasto	13.1167	149.5609	7.8777
	15.7867	181.4497	15.9305

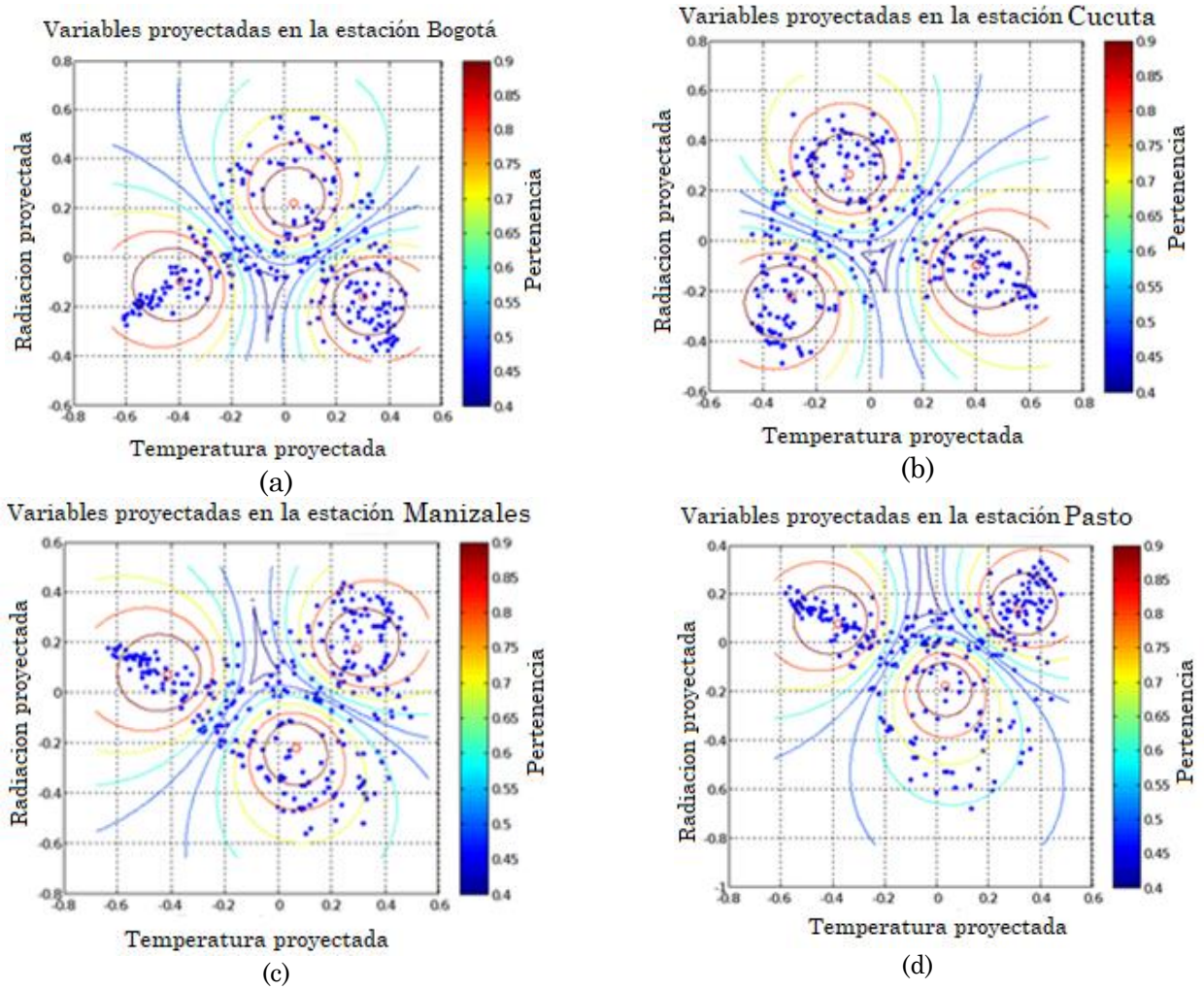

Fig. 7. Centroides proyectados en R^2 utilizando el algoritmo de visualización de PCA en las ciudades de: (a) Bogotá, (b) Cúcuta, (c) Manizales y (d) Pasto. Fuente: autores.

Tabla 13. Error relativo de reconstrucción obtenido de PCA. Fuente: autores.

Variable real	$Er(\%) = \frac{\|V_{lPCA} - V_{lFuzzy}\|}{\|V_{lFuzzy}\|} * 100$	**Etiqueta**
	0.9300	Mañana
Bogotá	0.7428	Medio día
	0.8917	Tarde
	6.0919	Mañana
Cúcuta	2.1289	Medio día
	1.6549	Tarde
	6.3867	Mañana
Manizales	2.0716	Medio día
	1.1544	Tarde
	0.7116	Mañana
Pasto	0.1394	Medio día
	1.0027	Tarde

5. CONCLUSIONES

La alta incidencia de los fenómenos hidrológicos en la generación de energía de Colombia, tales como el Fenómeno del Niño, durante periodos secos, muestra la necesidad apremiante de diversificar la matriz energética del país. Para ello, el gobierno nacional se ha comprometido con aumentar el uso de las FNCER entre el 13% y el 18% para el año 2031. Esto representa una oportunidad para la ejecución de proyectos de plantas de energía solar, cuya factibilidad podría estimarse a través de estudios similares al desarrollado en esta investigación. Así mismo, las principales conclusiones se muestran a continuación:

En este trabajo se encontró un potencial fotovoltaico por encima de los porcentajes recomendados en diferentes investigaciones. Los algoritmos K-means y Fuzzy C-means, proporcionaron factores de planta entre un 27% y un 36%, tal como se muestra en las tablas 4 y 9. Muchos autores coinciden en que un valor aceptable para este tipo de generación debe estar entre el 11% y el 24%, y ninguna tecnología con factor de capacidad menor al 10% tiene sentido de ser aplicada.

Debido a las inconsistencias que presenta la base de datos original, fue necesario un cuidadoso trabajo de limpieza en Excel inicialmente, el cual consistió en filtrar, ordenar y eliminar datos atípicos, de tal manera que facilitara su procesamiento en MATLAB®, conservando intactos sus parámetros estadísticos, tanto de tendencia central como de dispersión.

La minería de datos es una herramienta muy útil en el procesamiento de grandes volúmenes de información. En este proyecto se procesaron archivos con más de 270.000 datos por cada estación, cuyo manejo sería dispendioso, implicando altos costos computacionales, sin el uso de los algoritmos de agrupación existentes.

La cantidad de conglomerados trabajados en los dos métodos se validaron con cuatro índices (S(i), SI, SC y XB), los cuales determinaron el número óptimo de centroides en cada caso. En algunos resultados, se presentaron diferentes valores, pero cantidad predominante fue de 3 como la mejor alternativa de agrupamientos. Por tal motivo, el criterio del investigador juega un papel fundamental a la hora de determinar dicho número, ya que debe de ser éste quien finalmente tome la decisión basándose en datos como: valores más frecuentes, descarte de datos atípicos y presentar coherencia en la información. Las etiquetas para las clases seleccionadas corresponden a los nombres de “Mañana”, “Medio día” y “Tarde”.

Por otro lado, se plantean como trabajos futuros de investigación y/o recomendaciones las siguientes:

La utilización de diferentes algoritmos de agrupamiento, tanto predictivos como descriptivos, y otros índices de validación como internos, externos y relativos para comparar con los resultados actuales.

Estudios de correlación y regresión entre las variables consideradas en este estudio, así como otras, tales como la humedad y presión atmosférica, para determinar su incidencia sobre el cálculo del Factor de planta.

Extender el estudio de factibilidad de implementación de los algoritmos desarrollados a otras FCNER tales como: eólica, biomasa, geotérmica, entre otras.

6. CONFLICTOS DE INTERÉS

No se presenta ningún conflicto de interés enmarcado en el desarrollo del artículo.

7. REFERENCIAS

[1] R. Martinez and E. Forero, “Estimation of energy efficiency in solar photovoltaic panels considering environmental variables,” *IOP*

Conf. Ser. Mater. Sci. Eng., vol. 437, no. 1, p. 012008, Oct. 2018. https://doi.org/10.1088/1757-899X/437/1/012008

[2] E. J. Hernández-Leal, N. D. Duque-Méndez, and J. Moreno-Cadavid, "Big Data: una exploración de investigaciones, tecnologicas y casos de aplicación," *TecnoLógicas*, vol. 20, no. 39, pp. 17–24, Aug 2017. https://doi.org/10.22430/22565337.685

[3] M. Valencia, "Crisis energética en Colombia," *Tecnol. Investig. Y Acad.*, vol. 4, no. 2, pp. 74–81, Dec 2016. https://revistas.udistrital.edu.co/index.php/tia/article/view/10411/pdf

[4] G. J. López, I. A. Isaac, J. W. González, and H. A. Cardona, "Integración de energías renovables (solar fotovoltaica) en campus UPB laureles-micro red inteligente," *Investig. Apl.*, vol. 8, no. 2, pp. 152–159, Dec. 2014.

[5] L. González Polanco and G. Pérez Betancourt, "La minería de datos especiales y su aplicación en los estudios de salud y epidemiología," *Rev. Cuba. Inf. en Ciencias la Salud*, vol. 24, no. 4, pp. 482–489, 2013.

[6] B. Balasko, J. Abonyi, and B. Feil, "Fuzzy clustering and data analysis toolbox for use with matlab," Veszprem, Hungary. 2005. https://pdfs.semanticscholar.org/72f6/b22f6db1c2c0c47d8e6ead009b8c4c42bad9.pdf

[7] "Con la Política de Pago por Servicios Ambientales se da vía libre a los incentivos económicos para la conservación. " . 2017. Ministerio de Ambiente y Desarrollo Sostenible. 11 de Septiembre de 2019 http://www.minambiente.gov.co/index.php/noticias/3025-con-la-politica-de-pago-por-servicios-ambientales-se-da-via-libre-a-los-incentivos-economicos-para-la-conservacion.

[8] Procolombia, "Electric Power in Colombia Power Generation," Procolombia, 2015. http://www.energynet.co.uk/webfm_send/1839

[9] H. Grossi Gallegos, A. Roberti, G. Renzini, and V. Sierra, "Algunos comentarios sobre el modelo de Suehrcke y su aplicación en Argentina," *Av. en Energías Renov. y Medio Ambient.*, vol. 7, no. 2, pp. 11.01-11.05, Nov. 2003.

[10] D. F. Vallejo, "Clustering de Documentos con Restricciones de Tamaño," Universitat Politècnica de València, 2016.

[11] C. O. Solorio and D. P. Huertas, "Estimación de la radiación global para la República Mexicana (primera aproximación)," *Rev. Geogr. Agrícola*, pp. 77–84. https://chapingo.mx/revistas/revistas/articulos/doc/rga-1707.pdf

[12] Murcia, Humberto Rodríguez. "Desarrollo de la energía solar en Colombia y sus perspectivas." *Revista de ingenier*ía,pp 83-89, Jan. 2008. http://www.scielo.org.co/pdf/ring/n28/n28a12.pdf

[13] I. Zamarbide Ducun, "Predicción de radiación solar a corto y medio plazo," Universidad Publica de Navarra, Thesis de maestría, 2014. https://academica-e.unavarra.es/handle/2454/12164

[14] S. Guevara Vásquez, "Estimación de la radiación solar," CEPIS, 2003. http://www.bvsde.paho.org/bvsacd/cosude/xxii.pdf

[15] J. A. Duffie and W. A. Beckman, *Solar engineering of thermal processes*. John Wiley & Sons, 2013.

[16] L. Herrera, A. Miranda, E. I. Arango-Zuluaga, C. A. Ramos-Paja, and D. González-Montoya, "Dimensionamiento de sistemas de generación fotovoltaicos localizados en la ciudad de Medellín," *TecnoLógicas*, pp. 289–301, 2013. https://doi.org/10.22430/22565337.333

[17] B. J. Restrepo-Cuestas, A. Trejos-Grisales, and C. A. Ramos-Paja, "Modeling of PV Systems Based on Inflection Points Technique Considering Reverse Mode," *TecnoLógicas*, pp. 237–248, 2013. https://doi.org/10.22430/22565337.353

[18] J. D. Bastidas-Rodríguez, C. A. Ramos-Paja, and L. A. Trejos-Grisales, "Modelo Matemático de Arreglos Fotovoltaicos en Puente-Vinculado Operando bajo Condiciones Irregulares," *TecnoLógicas*, p. 223, Nov. 2013. https://doi.org/10.22430/22565337.361

[19] L. M. Pomares, "Análisis y predicción de series temporales de irradiancia solar global mediante modelos estadisticos," Thesis Doctoral, Universidad Complutense de Madrid, 2012.

[20] Y. Maldonado, G. Roncancio, and J. D. S. Saavedra, "Evaluación del potencial de energía solar en Santander, Colombia.," *Rev. Prospect.*, vol. 17, no. 2, pp. 7–12, Dec. 2019. https://doi.org/10.15665/rp.v17i2.1645

[21] "Tecnoglass," *Tecnoglass prevé invertir este año US$20 millones más en generación de energía solar*, 2017.

[22] E. L. Guzmán, "Métricas para la validación de Clustering," Elizabeth León Guzmán, 2016. https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

[23] C. Validated, "How to interpret mean of Silhouette plot?," *Cross Validated*, 2011. https://stats.stackexchange.com/questions/10540/how-to-interpret-mean-of-silhouette-plot/44653#44

[24] S. Villazana, F. Arteaga, C. Seijas, and O. Rodriguez, "Estudio comparativo entre algoritmos de agrupamiento basado en svm y c-medios difuso aplicados a señales electrocardiográficas arritmicas," *Rev. Ing. UC*, vol. 19, no. 1, pp. 16–24, Apr. 2012.

[25] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, 1991. https://doi.org/10.1109/34.85677

[26] B. H. Vergara, "Pago por potencia firme a centrales de generación Eólica," Thesis pregrado, Universidad de Chile, 2006. http://hrudnick.sitios.ing.uc.cl/paperspdf/HerreraB.pdf

[27] Díez J. L., Navarro J. L., and Sala A., "Algoritmos de clustering en la identificación de modelos borrosos," *Revista iberoameroamericana de Automática. Barcelona, Spain*, no. 2. pp. 32-41. Oct. 2010. https://polipapers.upv.es/index.php/RIAI/article/view/8010

[28] J. L. V. Villardón, "Análisis de componentes principales," *Cataluña UOC, Dep. Estadística*, vol. 32, 2002. http://benjamindespensa.tripod.com/spss/ACP.pdf

[29] Microsoft, "Métodos de discretización (minería de datos)." 2017. https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/discretization-methods-data-mining?view=sql-server-2014